Istraživanje tržišta za platformu za umjetnu inteligenciju

1. Opis projekta

Svrha ovog istraživanja tržišta je prikupljanje informacija o dostupnim hardverskim i softverskim rješenjima koja podržavaju izgradnju i upravljanje AI platformom namijenjenom izvođenju visokoperformantnih zadataka poput strojnog učenja (ML), dubokog učenja, znanstvene vizualizacije te rada s velikim jezičnim modelima (LLM).

Cilj je dobiti ponude za cjelokupno rješenje koje uključuje:

- Hardversku infrastrukturu
- Softversku platformu za upravljanje resursima
- GPU nadzor i upravljanje
- Skalabilnu i sigurnu administraciju korisnika i projekata
- Mogućnost daljnje nadogradnje platforme

2. Tehnički zahtjevi

2.1. Hardverski zahtjevi

Poslužitelj – Visokoperformantni za AI/ML

- Minimalno 2 procesora, svaki s najmanje 64 jezgre / 128 niti, minimalna frekvencija 3.2 GHz, Turbo, minimalno 256 MB Cache

- Minimalno 2.3 TB DDR5 memorije, konfiguracija u 24 modula po 96 GB, 6400MT/s, Dual Rank

- Minimalno 2 x 3.84 TB NVMe SSD diskova, Gen5, hot-swap

- Minimalno 8 grafičkih procesorskih jedinica (GPU), svaka s:

  - Minimalno 141 GB HBM3e memorije

  - PCIe Gen4 sučeljem

  - Potrošnjom od minimalno 450W

  - Podrškom za 4-way NVLink povezivanje

  - Fizički format: pasivno hlađena, double-wide, full height

  - Memorijska propusnost minimalno 4.8 TB/s

  - Minimalne FP16 (Tensor Core) performanse: 4.0 PFLOPS po jedinici

- Minimalno 2 x 1 Gb Ethernet + 2 x 10/25 GbE Ethernet priključci, s uključenim SFP28 modulima

- Potpuno redundantno (4+4) hot-plug napajanje, ukupno minimalno 3200W Titanium, podrška samo za 230-240V

- Daljinsko upravljanje kompatibilno s IPMI 2.0 i Redfish API standardima

- Maksimalna visina 4U za ugradnju u standardni 19'' poslužiteljski ormar

- Minimalno 5 godina jamstva

Poslužitelj – Osnovna konfiguracija

- Minimalno 1 procesor, 16 jezgri / 32 niti, minimalno 3.0 GHz, 64 MB Cache

- Minimalno 32 GB DDR5 5600MT/s, konfiguracija u 2 x 16 GB

- Minimalno 4 x 1.92 TB SSD SAS diskova, hot-swap, RAID 5 podrška, brzina do 24 Gbps

- Minimalno 2 x 1 GbE + 2 x 10/25 GbE priključci s uključenim modulima

- Dualno redundantno napajanje (1+1), minimalno 700W Titanium, 200-240V

- Daljinsko upravljanje putem sustava kompatibilnog s IPMI 2.0 i HTML5 konzolom

- Maksimalna visina 1U za ugradnju u standardni 19'' poslužiteljski ormar

- Minimalno 5 godina jamstva

2.2. Softverski zahtjevi

Platforma za upravljanje računalnim resursima

- Podrška za Proxmox VE ili ekvivalentnu virtualizacijsku platformu temeljenu na KVM tehnologiji

- Omogućeno kreiranje, brisanje i upravljanje VM-ovima

- Integracija alata za provisioning i nadzor (API pristup za automatsko pokretanje, gašenje, migriranje sustava)

- Centralizirani sustav autentifikacije i autorizacije (2FA, minimalni password policy, OAuth)

- Implementacija REST/GraphQL API-ja za upravljanje svim funkcionalnostima

- Prilagođeno web sučelje za krajnje korisnike (vanjske suradnike) s ograničenim pristupom uz obaveznu prijavu

Administrativna aplikacija

- Upravljanje korisnicima i timovima (kreiranje, uređivanje, brisanje korisničkih privilegija)

- Upravljanje zahtjevima za resurse (odobravanje, odbijanje)

- Evidencija svih aktivnih virtualnih instanci i bare-metal poslužitelja

- Generiranje izvještaja o potrošnji resursa (GPU vrijeme, CPU vrijeme, diskovni prostor, električna energija) u CSV i PDF formatima

- Audit log svih korisničkih aktivnosti

- Implementacija sigurnosnih mehanizama (2FA, GDPR usklađenost, anonimizacija podataka)

Tehnički nadzor GPU jedinica

- Softverski sustav mora omogućiti kontinuirano praćenje tehničkog stanja GPU jedinica

- Praćenje temperatura, opterećenja procesora i memorije, bilježenje ECC error-a

- Health checks i generiranje alarma u slučaju pregrijavanja, kvara ili nepravilnosti

- Integracija s nadzornim sustavima za pravovremenu obavijest administraciji

3. Pitanja za istraživanje tržišta

1. Predložite cjelovitu arhitekturu (hardver + softver) koja ispunjava navedene zahtjeve.

2. Objasnite skalabilnost predloženog rješenja (na koji način se platforma može nadograđivati u budućnosti).

3. Koji je očekivani vremenski okvir za isporuku, instalaciju i inicijalnu konfiguraciju predloženog rješenja?

4. Objasnite upravljanje GPU-ovima i AI workloadovima (nadgledanje, optimizacija, alarmiranje).

5. Predložite softverski sustav za administraciju korisnika, resursa i AI instanci.

6. Predložite način praćenja potrošnje GPU i CPU resursa po korisniku i projektu.

7. Objasnite podršku za izolaciju projekata (odvojeni tenants, sigurnosne domene).

8. Dostavite okvirnu cijenu za predložno rješenje (hardver + softver + osnovna instalacija).

9. Dostavite okvirne troškove podrške i održavanja za razdoblje od 5 godina.